Published online 8 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D983–D986***
doi:10.1093/nar/gkn709

AthaMap, integrating transcriptional and post-transcriptional data

Lorenz Bülow[1], Stefan Engelmann[2], Martin Schindler[2] and Reinhard Hehl[1,*]

[1]Institut für Genetik,Technische Universität Braunschweig, Spielmannstr. 7, D-38106 Braunschweig and [2]Software Systems Engineering Institute, Technische Universität Braunschweig, Mühlenpfordtstr. 23, D-38106 Braunschweig, Germany

Received September 1, 2008; Revised and Accepted September 29, 2008

ABSTRACT

The AthaMap database generates a map of predicted transcription factor binding sites (TFBS) for the whole *Arabidopsis thaliana* genome. AthaMap has now been extended to include data on post-transcriptional regulation. A total of 403 173 genomic positions of small RNAs have been mapped in the *A. thaliana* genome. These identify 5772 putative post-transcriptionally regulated target genes. AthaMap tools have been modified to improve the identification of common TFBS in co-regulated genes by subtracting post-transcriptionally regulated genes from such analyses. Furthermore, AthaMap was updated to the TAIR7 genome annotation, a graphic display of gene analysis results was implemented, and the TFBS data content was increased. AthaMap is freely available at http://www.athamap.de/.

INTRODUCTION

A large number of different databases are available for database-assisted gene-expression analysis (1). The first level of gene-expression regulation is transcription which is controlled by the synchronized binding of transcription factors (TFs) to adjacent *cis*-regulatory sequences. The bioinformatic identification of *cis*-regulatory sequences is an important tool to predict target genes of specific TFs (2). Towards these ends, the AthaMap database was developed. AthaMap is a database that generates a genome-wide map of predicted transcription factor binding sites (TFBS) and *cis*-regulatory elements for *Arabidopsis thaliana* (3,4). Compared to similar databases such as AGRIS, Athena and ATTED-II (5–8), AthaMap covers the whole-genome sequence and includes predicted TFBS that were identified with positional weight matrices. Recently, plant-related contents of the transcription and promoter databases TRANSFAC and TRANSPRO (9,10) were integrated with plant proteome and pathway data to the platform BKL Plant (BIOBASE Knowledge library). This was combined with the previously reported ExPlain tool that screens promoter regions with positional weight matrices for TFBS and evaluates results using the 'Composite Module Analyst' (CMA) as core component (11,12). This commercial product integrates promoter and pathway analysis of gene-expression data (BIOBASE, Wolfenbüttel, Germany).

In contrast, AthaMap is in the public domain and provides online tools to display TFBS in user-selected genes or at specific genomic positions (3). The detection of combinatorial elements and their target genes allows the prediction of co-regulated genes (13). The gene analysis function detects common TFBS in user-provided genes (14). A short user manual has been published recently (15) and all tools are explained on the 'Description' page on the AthaMap website as well. AthaMap has been linked with PathoPlant, a database on plant–pathogen interactions (16). *Arabidopsis thaliana* microarray experiments in PathoPlant can be screened for co-regulated genes that respond to up to three different stimuli (17). A list of co-regulated genes can directly be exported to AthaMap for identification of common TFBS. However, not all differentially expressed genes are transcriptionally regulated (18). One important factor for post-transcriptional regulation is the expression of small RNAs such as miRNA, siRNA and ta-siRNA (19). Although there are distinct pathways to generate these types of small RNAs, the resulting molecules are very similar in size and represent the small RNA transcriptome of the organism (20). Using a massive parallel sequencing approach, small transcriptome data became available for seedlings and inflorescence tissue of *A. thaliana* (21). The genome-wide nature of AthaMap and the availability of small RNA data provide a unique opportunity to combine transcriptional and post-transcriptional data in a single database. This may add significantly to the quality of *cis*-regulatory sequence identification involved in transcriptional regulation.

*To whom correspondence should be addressed. Tel: +49 531 391 5772; Fax: +49 531 391 5765; Email: r.hehl@tu-braunschweig.de

ANNOTATION OF GENOMIC POSITIONS OF SMALL RNAS

Sequence signatures (17-mers) derived from a small RNA transcriptome analysis of *A. thaliana* inflorescence tissue and seedlings were used for genomic screenings (21). The complete lists of screening sequences (Accession numbers GSM65747 and GSM65750) were downloaded from NCBI's Gene Expression Omnibus (GEO) repository (22). Genomic positions were determined by using a Perl script that screens for occurrences of perfect matches of all 109 590 small RNA 17-mer screening sequences within the five chromosomes of *A. thaliana*. Absolute positions and orientation of small RNA matches from inflorescence tissue and seedlings were annotated to AthaMap resulting in a total of 403 173 genomic matches. For screening sequences yielding more than one genomic match, corresponding loci were determined. A total of 5772 genes were predicted to be post-transcriptionally regulated by small RNAs since their transcribed regions are targets of at least one small RNA in antisense orientation. A text file with the genome identifiers of the 5772 predicted target genes of small RNAs can be downloaded on the documentation page at AthaMap.

Genomic positions of small RNAs are displayed in AthaMap analogous to TFBSs and are symbolized as xxxxx>. The arrow head gives the orientation of the small RNA. A tool tip box appears when moving over the arrow indicating the absolute genomic position and screening library of the small RNA. Selecting the name adjacent to this symbol will open a new window giving additional information. Figure 1 shows a partial screen shot of position 11 911 on chromosome 1 with a small RNA from the inflorescence library, the tool tip box and the associated pop-up window. This new window shows the screening sequence, corresponding genomic positions for this particular small RNA and the reference.

Putative post-transcriptionally regulated genes are identified within the Colocalization and Gene Analysis functions. These genes are tagged on the result pages with an italicized genome identifier. They can be subtracted in the Colocalization and Gene Analysis functions by activating the checkbox 'exclude genes regulated by smallRNA' in order to restrict the analyses exclusively to transcriptionally regulated genes.

UPDATE TO TAIR7

The recent publication of the TAIR7 *A. thaliana* genome release motivated the implementation of this genome annotation into AthaMap (23). The annotation of the gene structure is based on five chromosomal XML flatfiles downloaded from the TAIR web site (release 7). These files were parsed using a Perl script and positional information for 5′- and 3′-UTRs, exons and introns were annotated to AthaMap. These regions are displayed in AthaMap with a colour code similar to the one used by TAIR. Due to the significantly increased number of genes with annotated transcription start site (TSS) in TAIR7, the Gene Analysis and Colocalization functions of AthaMap have been changed to show positions of TFBS relative to TSS of the nearest gene. This applies to 23 222 (73.1%) genes while for the remaining 8540 (26.9%) genes results are still displayed relative to the translation start site. In earlier versions of AthaMap, all positions were shown relative to translation start sites as point of reference. Compared to TAIR5 the previous version annotated to AthaMap, the nucleotide sequence of the *A. thaliana* genome in TAIR7 was not changed. Therefore, the positional information of all previously determined TFBS remained constant, except for TATA-boxes. Because of the larger number of genes with an annotated TSS, the number of annotated TATA boxes decreased from 16 277

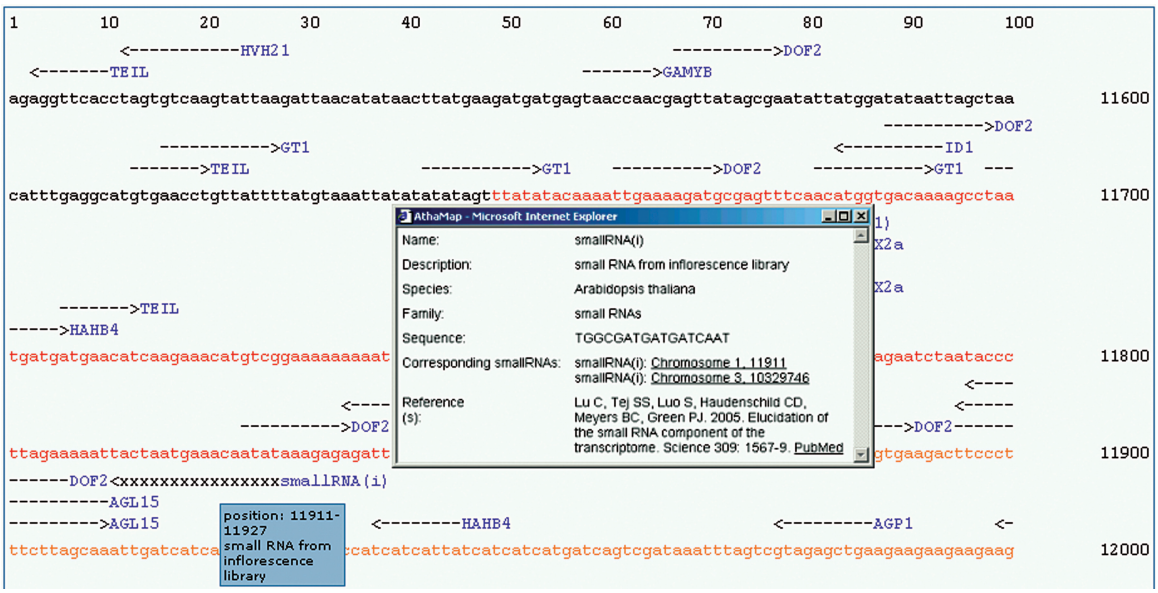

Figure 1. Small RNA binding sites in the *Arabidopsis thaliana* genome. Partial screen shot of the sequence display window with a small RNA binding site at position 11 911 on chromosome 1. The tool tip box indicates the absolute genomic position and screening library. A pop-up window with additional information on the small RNA is also shown.

(13) to currently 15955. The number of TATA boxes decreased because for genes lacking a TSS a larger upstream region was screened for putative TATA boxes than for genes with an annotated TSS (3). Therefore, the lower number of TATA boxes results from elimination of false positives.

GRAPHIC DISPLAY OF GENE ANALYSIS RESULTS

The Gene Analysis function of AthaMap generates long lists with positional information on TFBSs in all genes investigated (14). Although overviews or summaries of the data can be displayed, the positional information is difficult to perceive. Therefore, a graphic display of TFBS in the analysed gene region was implemented that enables easy comparison between genes and visual identification of common binding site patterns. Every TF family as well as the small RNAs and combinatorial elements are identified with a different colour and their display can be selected individually. Figure 2 shows the web interface with the buttons to select the TF families and a graphic display of TFBS for selected TF family members in the *Arabidopsis* genes At2g42530 and At2g42540. Also shown is a tool tip box that opens when the mouse pointer moves over the colour-coded TFBS. The tool tip box gives additional information for the TF that identified this particular TFBS. Factor (RAV1) and factor family (AP2/EREBP) are identified as well as the position relative to the TSS (−70). For TFBS identified with positional weight matrices, threshold score, maximum score and score of the binding site are given (3).

DATA INCREASE

Recently published binding sites for the *Arabidopsis* TFs TAC1, RAP2.2 and MYB98 were annotated to AthaMap (24–26). These factors belong to the C2H2(Zn), AP2/EREBP and MYB TF families. Detection and annotation

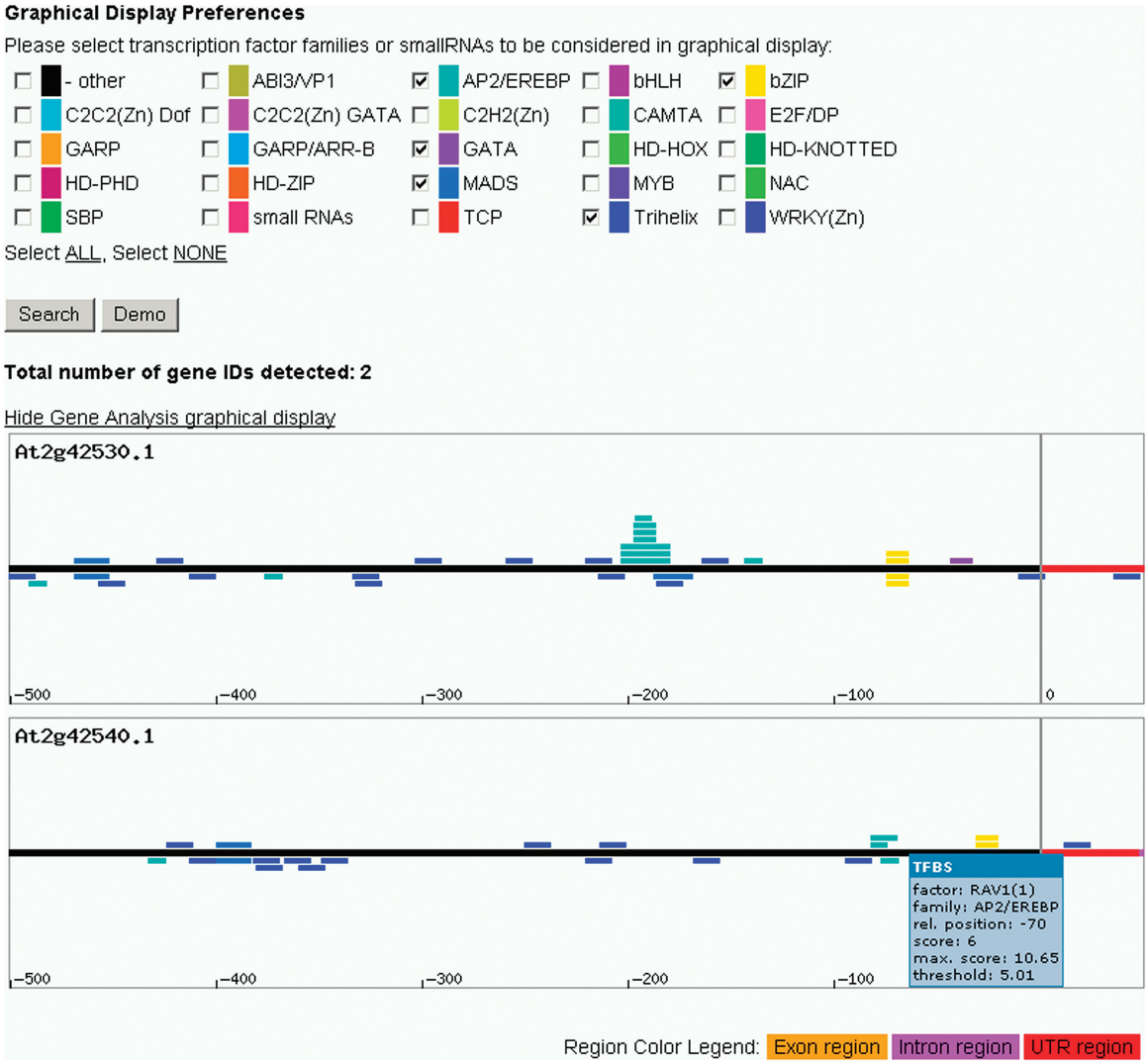

Figure 2. Graphic display of transcription factor and small RNA binding sites. Partial screen shot of the gene analysis tool with the checkboxes for TF families included in a graphic display and the graphic display of the upstream region of the genes At2g42530 and At2g42540. A tool tip box with additional information on one of the TFBS is also shown.

of single binding sites was done as described earlier (4). Binding sites for two TFs for which positional weight matrices could be generated were annotated as well. These are the factors STF1 and SPL1 which belong to the bZIP and SBP TF families (27,28). Detection and annotation of matrix-based binding sites was done as described earlier (3). AthaMap now harbours 9 998 736 predicted TFBSs.

ACKNOWLEDGEMENTS

We would like to thank Anne-Kareen Blechert for help implementing the TAIR7 genome annotation and for TFBS screenings.

FUNDING

German Federal Ministry for Education and Research through GABI-ADVANCIS (BMBF 0315037B). Funding for open access charge: Technical University of Braunschweig.

Conflict of interest statement. None declared.

REFERENCES

1. Hehl,R. and Bülow,L. (2008) Internet resources for gene expression analysis in *Arabidopsis thaliana*. *Curr. Genomics*, **9**, 375–380.
2. Hehl,R. and Wingender,E. (2001) Database-assisted promoter analysis. *Trends in Plant Sci.*, **6**, 251–255.
3. Steffens,N.O., Galuschka,C., Schindler,M., Bülow,L. and Hehl,R. (2004) AthaMap: an online resource for in silico transcription factor binding sites in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.*, **32**, D368–D372.
4. Bülow,L., Steffens,N.O., Galuschka,C., Schindler,M. and Hehl,R. (2006) AthaMap: from *in silico* data to real transcription factor binding sites. *In Silico Biol.*, **6**, 0023.
5. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
6. O'Connor,T.R., Dyreson,C. and Wyrick,J.J. (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, **21**, 4411–4413.
7. Palaniswamy,S.K., James,S., Sun,H., Lamb,R.S., Davuluri,R.V. and Grotewold,E. (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol.*, **140**, 818–829.
8. Obayashi,T., Kinoshita,K., Nakai,K., Shibaoka,M., Hayashi,S., Saeki,M., Shibata,D., Saito,K. and Ohta,H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.
9. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
10. Chen,X., Wu,J.M., Hornischer,K., Kel,A. and Wingender,E. (2006) TiProD: the Tissue-specific Promoter Database. *Nucleic Acids Res.*, **34**, D104–D107.
11. Kel,A., Voss,N., Jauregui,R., Kel-Margoulis,O. and Wingender,E. (2006) Beyond microarrays: finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics*, **7(Suppl 2)**, S13.
12. Kel,A., Konovalova,T., Waleev,T., Cheremushkin,E., Kel-Margoulis,O. and Wingender,E. (2006) Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics*, **22**, 1190–1197.
13. Steffens,N.O., Galuschka,C., Schindler,M., Bülow,L. and Hehl,R. (2005) AthaMap web tools for database-assisted identification of combinatorial *cis*-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*. *Nucleic Acids Res.*, **33**, W397–W402.
14. Galuschka,C., Schindler,M., Bülow,L. and Hehl,R. (2007) AthaMap web-tools for the analysis and identification of co-regulated genes. *Nucleic Acids Res.*, **35**, D857–D862.
15. Hehl,R. (2008) In Kahl,G. and Meksem,K. (eds), *The Handbook of Plant Functional Genomics: Concepts and Protocols*. Wiley and Sons Ltd, Weinheim, Germany, pp. 337–346.
16. Bülow,L., Schindler,M., Choi,C. and Hehl,R. (2004) PathoPlant®: a database on plant-pathogen interactions. *In Silico Biol.*, **4**, 529–536.
17. Bülow,L., Schindler,M. and Hehl,R. (2007) PathoPlant®: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids Res.*, **35**, D841–D845.
18. Cheadle,C., Fan,J., Cho-Chung,Y.S., Werner,T., Ray,J., Do,L., Gorospe,M. and Becker,K.G. (2005) Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC Genomics*, **6**, 75.
19. Jones-Rhoades,M.W., Bartel,D.P. and Bartel,B. (2006) MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.
20. Vaucheret,H. (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev.*, **20**, 759–771.
21. Lu,C., Tej,S.S., Luo,S., Haudenschild,C.D., Meyers,B.C. and Green,P.J. (2005) Elucidation of the small RNA component of the transcriptome. *Science*, **309**, 1567–1569.
22. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
23. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
24. Ren,S., Mandadi,K.K., Boedeker,A.L., Rathore,K.S. and McKnight,T.D. (2007) Regulation of telomerase in Arabidopsis by BT2, an apparent target of TELOMERASE ACTIVATOR1. *Plant Cell*, **19**, 23–31.
25. Welsch,R., Maass,D., Voegel,T., Dellapenna,D. and Beyer,P. (2007) Transcription factor RAP2.2 and its interacting partner SINAT2: stable elements in the carotenogenesis of Arabidopsis leaves. *Plant Physiol.*, **145**, 1073–1085.
26. Punwani,J.A., Rabiger,D.S. and Drews,G.N. (2007) MYB98 positively regulates a battery of synergid-expressed genes encoding filiform apparatus localized proteins. *Plant Cell*, **19**, 2557–2568.
27. Song,Y.H., Yoo,C.M., Hong,A.P., Kim,S.H., Jeong,H.J., Shin,S.Y., Kim,H.J., Yun,D.J., Lim,C.O., Bahk,J.D. *et al.* (2008) DNA-binding study identifies C-box and hybrid C/G-box or C/A-box motifs as high-affinity binding sites for STF1 and LONG HYPOCOTYL5 proteins. *Plant Physiol.*, **146**, 1862–1877.
28. Liang,X., Nazarenus,T.J. and Stone,J.M. (2008) Identification of a consensus DNA-binding site for the *Arabidopsis thaliana* SBP domain transcription factor, AtSPL14, and binding kinetics by surface plasmon resonance. *Biochemistry*, **47**, 3645–3653.